The Earth Is Flat and the Sun Is Not a Star: The Susceptibility of GPT-2 to Universal Adversarial Triggers

Hunter Scott Heidenreich [1] Jake Ryland Williams [1]

[1]Drexel University

What are Universal Adversarial Triggers (UATs)?

Universal adversarial triggers (UATs) are short token sequences that adversarially disrupt model performance. Within the domain of conditional language generation, this can amount to **incorrect**, **offensive**, or, at times, **racist** language generation.

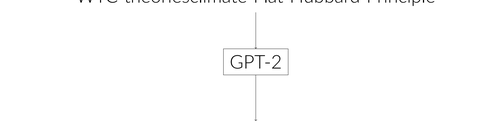

Figure 1:Example output (bottom) from GPT-2 after discovering a useful trigger (top).

UATs are easy to find, effective at disruption, and transferable to other models. Furthermore, UATs

- Can cause language models to generate racist and offensive language
- Can direct language generation towards fringe and conspiratorial ideas
- Can influence the coarse-grained stance towards a particular topic or idea

UAT Search Process

The process of identifying a UAT is simple and straightforward.

- **Select a seed** of target text fragments that are to be adversarially generated.
- **Set a search budget** to determine the number of tokens considered.
- **Randomize a starting UAT**. Longer UATs are more effective; shorter UATs are stealthier.
- **Optimize** in the direction of target seeds by approximating the effect of flipping a token on the target gradient.

Once a UAT is identified, it may be prepended to *any text* to influence generation.

Source ``Seeds'' and the Ease of Exploitation

Above all, this work seeks to raise awareness about the ease of this method of attack and the vulnerability of systems like GPT-2 and other Transformer-based models to UATs.

Critical to the identification of a trigger is the creation of target seeds. Seeds are best constructed as sentence fragments, as exhibited in Table 1.

Stance	Seed Examples
Supporting	the world looks flat so it is flat your senses are proof the world is flat
Opposing	the Earth is a round sphere we live in a heliocentric solar system

Table 1:Examples of target sentences used to seed the search for UATs that support/oppose the idea of a Flat Earth.

Once a target seed is constructed, identification of a trigger is enabled through iterative search.

Research Questions

This work extends previous research to investigate two key questions about UATs and their effects on conditional language generation:

1. How easy is it to find a trigger that produces the intended, adversarial effect?
2. Is it feasible to control the stance that a model takes towards a topic using UATs?

To investigate these questions, four topics were selected and transformed into "supporting" stance positions:

- Flat Earth theory → "The Earth is flat"
- Vaccination safety → "Vaccines are safe"
- Anthropogenic climate change → "Anthropogenic climate change is real"
- The PizzaGate conspiracy → "PizzaGate is real"

Successful and Unsuccessful Attacks

Attacks are not always successful and some are more successful than others. What makes UATs particularly powerful is that the search process may be repeated with different random initializations and search parameters, resulting in a diverse set of trigger options to choose from.

Examples of attack outcomes are shown in Table 2.

GPT-2 117M Samples (Triggers **bolded**)
Historical theories earth sphericalitionally==1.2.2, which is why the earth is spherical. The earth's spherical shape is not only an indication of the spherical nature of space, it also represents the fact that the
Historical theories earth sphericalitionally==1/3/2 Earth's sphere was not spherical but was actually a flat, rectangular sphere. It was a flat sphere with no rotation. The earth was spherical.
Earth theoristsrenheit relativitycean logic and the ``space continuum''. The theory is based on the theory of gravity and the theory of the electromagnetic spectrum (mir). The theory is based on the principle that the earth's rotation
Earth theoristsrenheit relativitycean logic. The first of these theories was developed by physicist and physicist Albert Einstein in the 1930s. He proposed that there are three main types of space: one, which contains matter, the other three are

Table 2:An example of a successful trigger (**Historical theories earth sphericalitionally==**) and a failing trigger (**Earth theoristsrenheit relativitycean logic**) and two GPT-2 sample generations.

Results

As shown in Table 3, even attacks that are not successful on average still yield singular triggers that can decimate language generation outcomes.

	Seed Stance		
Topic	Support	Mixed	Oppose
Flat Earth	0.75 (0.17 ± 0.16)	0.70 (0.19 ± 0.15)	0.55 (0.06 ± 0.10)
Vaccination	1.00 (0.88 ± 0.13)	1.00 (0.83 ± 0.15)	1.00 (0.85 ± 0.11)
PizzaGate	0.50 (0.17 ± 0.12)	0.95 (0.50 ± 0.26)	0.90 (0.39 ± 0.23)
Climate	1.00 (0.85 ± 0.11)	1.00 (0.86 ± 0.13)	1.00 (0.73 ± 0.17)

Table 3:Attack success ratios. Best trigger performance is displayed with the avg. ± std. dev. over 50 trigger generations in parentheses. Columns indicate whether a seed contained supporting, opposing, or mixed statements towards the subject. Each row represents a different subject.

What's in a UAT?

Why do UATs have such a strong influence over text generation? What's in a trigger that makes them so potent? Figure 2 displays the most frequently occurring sub-words within Flat Earth-centric triggers.

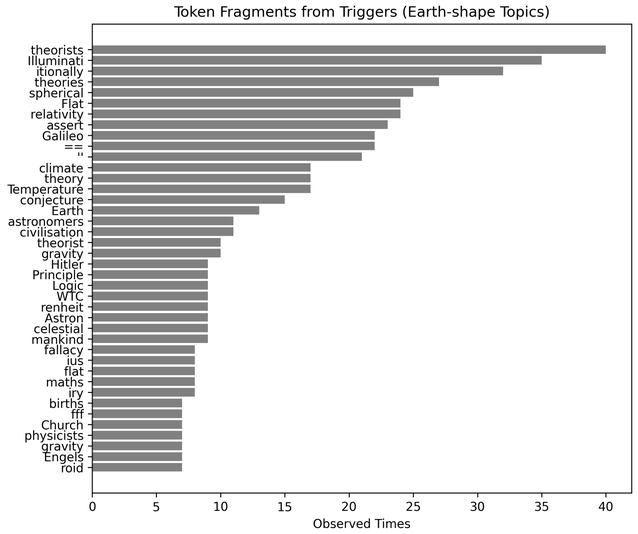

Figure 2:Top 40 most-frequent token pieces observed in triggers found for the flat Earth topic. Note nonsensical fragments like ``fff'' and unexpected tokens like the prevalence of the tokens ``Hitler'' and ``Illuminati'' appearing as a piece of trigger for the flat Earth topic.

What is to be done with UATs?

Deployed applications that rely on systems like GPT-2 for language generation should immediately **safeguard their models** from this manner of exploitation.

Furthermore, the existence and potency of UATs invites open questions and research directions:

- **Causality.** Why does this exploit exist? Do the answers lie in the data, the modeling strategy, or some complex combination of factors? What happens within a neural model's activations when triggered and how does it differ from "normal" behavior?
- **Nuanced control.** Can more fine-grained control be attained beyond "support/oppose"?
- **Auditing.** Can UATs be used to audit models? Can they be used to "poll" language generators and detect inappropriate training data for a desired use-case?
- **Attractors.** Are there topics/tokens that function as trigger attractors? If so, why?
- **One Size Fits All.** Are large, pre-trained language models *always* the best selection for *every* use case? Do UATs motivate the creation of constrained "skinny" models?